pennsylvania
OFFICE OF ADMINISTRATION

Information Technology Policy

Data Warehousing Policy

ITP Number ITP-INF004	Effective Date November 7, 2006
Category Information	Supersedes None
Contact RA-ITCentral@pa.gov	Scheduled Review May 2022

1. Purpose

This Information Technology Policy (ITP) establishes enterprise-wide standards and guidance for Data Warehousing.

2. Scope

This ITP applies to all offices, departments, boards, commissions and councils under the Governor's jurisdiction (hereinafter referred to as "agencies"). Agencies not under the Governor's jurisdiction are strongly encouraged to follow this ITP.

Third-party vendors, licensors, contractors or suppliers shall meet the policy requirements of the Commonwealth's ITPs that are applicable to the products and services provided to the Commonwealth.

3. Definitions

3.1 Administrative Data: The data used by a warehouse administrator to manage data in a data warehouse. Examples of Administrative Data are *user profiles* and *order history data.*

3.2 Aggregations: The process of consolidating data values into a single value. For example, sales data could be collected daily and then aggregated at the week and/or month level. The data can then be referred to as *aggregate data*. Aggregation is synonymous with *summarization*, and *aggregate data* is synonymous with *summary data*.

3.3 Analyst: A user who creates views for analytic interpretation of data, performs calculations, and distributes the resulting information in the form of reports.

3.4 Data Integration Technology: The consolidation and reconciliation of dispersed data maintained by agencies in multiple, heterogeneous systems for analytical purposes. Data can be accessed, extracted, moved, loaded, validated, and transformed.

3.5 Data Lake: A system or repository of data stored in its natural/raw format, usually object blobs or files. A data lake is usually a single store of data including raw copies of source system data, sensor data, social data, etc., and transformed data used for tasks such as reporting, visualization, advanced analytics and machine learning. A data lake can include structured data from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs) and binary data (images, audio, video)

3.6 Data Loading: Data Loading is the process of populating the data warehouse. Data Loading is provided by Database Management System (DBMS)-specific load processes, DBMS insert processes, and independent fast-load processes.

3.7 Data Mapping: Data Mapping is the assignment of a source data element to a target data element.

3.8 Data Mart: A data warehouse designed for a particular line of business, such as sales, marketing, or finance. In a dependent Data Mart, the data can be derived from an enterprise-wide data warehouse. In an independent Data Mart, data can be collected directly from sources.

3.9 Data Mining: Data Mining is the process of sifting through large amounts of data to produce data content relationships. It also refers to the technique by which a user utilizes software tools to look for particular patterns or trends. This technique can uncover future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. It is also known as *data surfing*.

3.10 Data Model: A Data Model is a logical map that represents the inherent properties of the data independent of software, hardware, or machine performance considerations. The Data Model shows data elements grouped into records, as well as the association around those records.

3.11 Data Warehousing: A process for building decision support systems and a knowledge-based application environment in support of both everyday tactical decision making and long-term business strategy. Data warehouses and data warehouse applications are designed primarily to support the decision-making process by providing the decision makers with access to accurate and consolidated information from a variety of sources.

3.12 Dimension Tables: Dimension Tables describe the business entities of an enterprise, represented as hierarchical, categorical information such as *time, departments, locations, and products*. Dimension Tables are sometimes called *lookup* or *reference tables.*

3.13 Extract, Transform and Load (ETL): Refers to the methods involved in accessing and manipulating source data and loading it into a data warehouse.

3.14 Fact: A value or measurement, which represents a Fact about the managed entity or system. Facts, as reported by the reporting entity, are said to be at raw level; Examples include sales, cost, and profit.

3.15 Fact Table: A table in a Star Schema that contains Facts. A Fact Table typically has two types of columns: Those that contain Facts (e.g., numbers), and those that are foreign keys to Dimension Tables. The primary key of a Fact Table is usually a composite key constructed with all its foreign keys.

3.16 Snowflake Schema: A Snowflake Schema is a set of tables comprised of a single, central Fact Table surrounded by normalized dimension hierarchies. Each dimension level is represented in a table. Snowflake Schemas implement dimensional data structures with fully normalized dimensions. Snowflake Schemas are an alternative to Star Schemas.

3.17 Star Schema: A Star Schema is a relational schema whose design represents a multidimensional Data Model. The Star Schema consists of one or more Fact Tables and one or more Dimension Tables related through foreign keys.

3. Objective

The primary objective of Data Warehousing is to collate information from disparate sources and place the information in a format conducive to the decision-making process. This objective necessitates a set of activities more complex than merely collecting data and reporting against it.

Data Warehousing requires both business and technical expertise and typically involves the following activities:

- Accurately identifying information to be placed in the data warehouse;
- Extracting, cleansing, aggregating, transforming, and validating the data to ensure accuracy and consistency;
- Identifying and prioritizing subject category areas to be included in the data warehouse;
- Managing the scope of each subject area implemented into the data warehouse on an iterative basis;
- Developing a scalable architecture to serve as the data warehouse's technical and application foundation and identifying and selecting the hardware/software/middleware components to implement it;
- Defining the correct level of consolidation to support business decision making;
- Establishing a refresh program that is consistent with business needs, timing, and cycles;
- Providing powerful and user-friendly tools at the desktop to access the data in the warehouse;
- Educating users about the business benefits made possible through Data Warehousing;
- Establishing data warehouse support mechanisms and training users to effectively utilize the desktop tools;
- Establishing processes for maintaining, enhancing, and ensuring the ongoing success and applicability of the data warehouse;
- Establishing governance processes for ensuring logging, auditing, security, and access controls associated with the data warehouse

4. Policy

Data Warehouse Technology Components:

Since Data Warehousing encompasses many technologies, it is not limited to one specialized area. Typically, the technical aspects of Data Warehousing are divided into the following areas:

- **ETL tools** address the process of extracting, transforming, and loading data from various agency application sources into the operational data store/data warehouse using either custom-developed utilities or existing marketplace products.

- **Data Warehouse repository tools** address products used for physical storage of data warehouse information. These tools range from products currently available in the marketplace; tools used in the Commonwealth; and those recommended for deployment.

- **Enterprise reporting tools** address the end-user reporting tools to be used to satisfy agency- specific reporting requirements.

- **Architecture standards** address the various architecture patterns and available modeling techniques; techniques currently used in the Commonwealth; and emerging modeling techniques.

- **Data exchange and security standards** address the information exchange mechanisms for horizontal and vertical information exchange (e.g., exchange of information within the Commonwealth, as well as with local, state and federal agencies). Security standards address authentication and access to the data warehouse for agency employees, and approval and audit processes by authorized agency members, subject to agency/enterprise security and privacy policies.

Warehousing requires combining both structured and unstructured data and/or where data needs to be stored in its natural/raw format (aka “data lake”) are outside the scope of this ITP.

Agencies are to utilize existing Data Warehousing solutions or build and implement a Data Warehousing solution when a business requirement necessitates reporting that summarizes or combines data from multiple sources.

Agencies are to leverage the Data Warehousing solution provided by Integrated Enterprise Systems (IES) for Enterprise Resource Planning (ERP) applications and/or applications that are associated with ERP data when the IES Data Warehousing solution meets agencies business requirements. ERP applications include financial, human resources, customer relationship management, supplier relationship management, platform life cycle management, supply chain management, and material management enterprise applications.

Data Warehousing standards are defined for Integrated Enterprise Systems in STD-INF004C *Data Warehousing Product Standards.*

The data warehouse is to enable access to centrally stored information to accommodate business reporting requirements.

The data warehouse is to contain a subset of information from operational systems optimized for data retrieval and reporting to support performance measurement against agency/enterprise goals and objectives.

Access to the data warehouse and the information within the data warehouse is to be based upon each user's job requirements and access level approved by the authorized agency officer, subject to agency/enterprise security and privacy policy.

To promote and maintain consistency across Commonwealth agencies, any data warehouse model is to follow ITP-INF003D *Core Citizen Data Model and Data Elements* for citizen-centric common data elements described in the citizen model. In addition, the data

warehouse model is to follow database standards referenced in ITP-INF001 *Database Management Systems*.

Data warehouse solutions are to enable data federation to support horizontal and vertical exchange (between agencies and potential future centralized Commonwealth-wide data warehouses).

Any data warehouse is to reside on hardware separate from operational and transaction-related systems, thereby mitigating potential performance issues with these systems.

Any data warehouse is to have an efficient extracting/harvesting process separate from operational and transaction-related systems in order to minimize the impact in either performance or availability of these systems.

Any custom development done for ETL is to adhere to existing Commonwealth standards.

Standard methods such as ANSI-SQL, Open Database Connectivity (ODBC), Java Database Connectivity (JDBC), and/or OData (Open Data Protocol) RESTful APIs will be used to access any data warehouse.

Due to privacy and security constraints, Data Warehousing solutions will physically operate on Commonwealth-approved infrastructure.

Data quality and accuracy are critical to establishing the integrity of the information and user confidence in the validity of the resulting output. Agencies are responsible for taking appropriate automated and manual measures to maintain a high degree of quality and accuracy of the information in the data warehouse.

Training requirements for each model of the data warehouse are to be met before a user will be granted access to data.

Agencies will determine the level of mission criticality of their data warehouse. The infrastructure and operational procedures necessary to support the data warehouse will be designed and implemented commensurate to the level of mission criticality of the data warehouse.

GEN-INF004A: *Introduction to Data Warehousing* provides an introduction to data warehousing technology.

BPD-INF004B: *Best Practice Approach to Data Warehousing* documents best practices in Data Warehousing.

STD-INF004C: *Data Warehousing Product Standards* provides guidance to agencies on the current standards and the status of other Data Warehousing solutions that are being used or being considered for use.

GEN-INF004D*: Data Warehousing Product Availability* provides information on the availability and licensing of current Data Warehousing product standards.

5. Responsibilities

5.1 Agencies shall comply with the requirements outlined in this ITP.

5.2 Third-party vendors, licensors, contractors, or suppliers creating custom applications on behalf of Commonwealth entities shall comply with the requirements as outlined in this ITP.

6. Related ITPs/Other References

Definitions of associated terms of this policy are published on the Office of Administration's public portal: http://www.oa.pa.gov/Policies/Pages/Glossary.aspx

Commonwealth policies, including Executive Orders, Management Directives, and IT Policies are published on the Office of Administration's public portal: http://www.oa.pa.gov/Policies/Pages/default.aspx

- Management Directive 205.34 Amended *Commonwealth of Pennsylvania Information Technology Acceptable Use Policy*
- ITP-ACC001 *Information Technology Digital Accessibility Policy*
- ITP-INF001 *Database Management Systems*
- BPD-INF003D *Core Citizen Data Model and Data Elements*
- GEN-INF004A *Introduction to Data Warehousing*
- BPD-INF004B *Best Practice Approach to Data Warehousing*
- STD-INF004C *Data Warehousing Product Standards*
- GEN-INF004D *Data Warehousing Product Availability*
- Process/Project DWH - Data Warehouse Process

7. Authority

Executive Order 2016-06 *Enterprise Information Technology Governance*

8. Publication Version Control

It is the Authorized User's responsibility to ensure they have the latest version of this publication, which appears on https://itcentral.pa.gov for Commonwealth personnel and on the Office of Administration public portal: http://www.oa.pa.gov/Policies/Pages/default.aspx. Questions regarding this publication are to be directed to RA-ITCentral@pa.gov.

9. Exemption from This Policy

In the event an agency chooses to seek an exemption from the guidance within this IT policy, a request for a policy waiver is to be submitted via the enterprise IT policy waiver process. Refer to ITP-BUS004 *IT Policy Waiver Review Process* for guidance.

This chart contains a history of this publication's revisions. Redline documents detail the revisions and are available to CWOPA users only.

Version	Date	Purpose of Revision	Redline Link

Original	11/7/2006	Base Document	N/A
Revision	03/23/2009	Added the use of IES Data Warehouse for ERP related applications.	
Revision	11/18/2010	ITP Refresh	
Revision	05/14/2021	ITP Refresh Added to ITP template Added Third-Party vendor to Scope and Responsibilities Updated Exemption Section	Revised IT Policy Redline <05/14/2021>